ITEA Journal 2009; 30: 251–257

Training and Testing Revisited: A Modest Proposal for Change

John K. Hawley, Ph.D. and Anna L. Mares
U.S. Army Research Laboratory, Fort Bliss, Texas

The present article is a follow-up to an earlier ITEA Journal article addressing the increasingly complex relationship between training and operational testing (see Hawley 2007). In the initial article, the lead author argued that effective test player (individual, crew, or unit) preparation is essential to valid operational testing. Inadequate preparation invariably results in a flawed test and undermines the validity of data essential to system improvement and acquisition decision making. The initial article outlined a set of pretest training actions that must occur if test players are to be properly prepared to participate in meaningful operational testing. These actions fell into three categories: (1) establishing a stable performance environment prior to testing, (2) pretest training conduct, and (3) pretest training evaluation. With respect to pretest training, Hawley (2007) concluded by asserting that test planners are faced with two choices: Plan and conduct adequate pretest training, or live with the consequences of a flawed test. The present discussion has an admittedly Army flavor, and many of our observations are taken from tests on Army systems, but the observations are generally applicable to other classes of systems and to other services as well.

Key words: Test player training; test validity; operational testing; human–machine systems; Patriot PDB-6; knowledge-based systems; alpha and beta testing; deployment standard.

Critics of the argument advanced in Hawley (2007) frequently respond that the training path that he proposes is not necessary and is not affordable in the increasingly cost-conscious defense acquisition environment. The author's reply to these criticisms is that a proper response to legitimate test constraints is not to ignore or downplay essential testing prerequisites and proceed as if test results are valid. That approach invites considerable outside scrutiny and criticism. If a valid test cannot be conducted within time and resource constraints, then the test's objectives must be simplified, or testing concepts revised in view of the resources likely to be available, balanced against requirements for valid testing—including adequate preparation of test players. Valid, in the present context, means that test results (1) accurately represent system performance capabilities and (2) reasonably generalize to a future operational setting. Testers must also bear in mind that in an *operational* test we are evaluating a *manned* system, and the manning component (e.g., the operators and maintainers) must be given consideration along with hardware and software capabilities.

The discussion to follow takes up where the first article left off and proposes several practical options for breaking out of the training–testing bind. We begin by reviewing the argument for adequate test player training as an essential prerequisite for valid testing. Next, recent research outlining training requirements for highly complex systems is reviewed, and the implications of this research for pretest training are discussed. These two sections define what must be done up front if test players are to be properly prepared to participate in operational testing. We also argue that these requirements cannot be ignored or traded away in the interests of time, schedule, or cost. The down-stream consequences for test validity can be fatal. We emphasize this point because of our recent experiences in Army operational testing where system evaluation has been seriously undermined by compromising on requirements for test player preparation. The final section outlines several steps that hold potential for lessening the growing impasse between adequate test

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **JUN 2009**	2. REPORT TYPE	3. DATES COVERED **00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE **Training and Testing Revisited: A Modest Proposal for Change**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Research Laboratory,Fort Bliss,TX,79908**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**7**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

player preparation and valid operational testing. The path out of the training–testing bind is conceptually straightforward, but might require a change in the way the Army views test player participation along the continuum ranging from developmental tests through full-scale operational tests.

The impact of inadequate test player preparation

We note in the previous paragraph that inadequate test player training compromises the validity of test results and thereby undermines the basis for acquisition decision making. The Government Accountability Office (GAO) has stated bluntly that inadequate test player preparation inevitably results in what is termed a "hollow" test (GAO 2000): The requirement to hold a testing "event" is satisfied, but the results do not advance system-related knowledge. Inadequate test player preparation effectively turns an operational test into little more than a technical demonstration. If the manned component is not able to provide reliable data, test results are compromised, and it is not possible to assess the system's fitness for use or honestly evaluate and refine usage concepts. Anyone familiar with system development is aware of the considerable gap between demonstrating a technical capability and deploying an operational system based on that capability.

In a hollow test, the most common form of compromise is *confounding* between test results and pretest proficiency levels. Confounding means that it is not possible to determine unambiguously whether observed test outcomes reflect materiel system capabilities and features, test player proficiency, or some interaction between the two. Posttest analysts cannot disentangle observed system performance from test player proficiency, regardless of the sophistication of the analytical methods used. All claims to the contrary notwithstanding, posttest analyses cannot compensate for an intrinsically flawed test. Testers would like to generalize from the test setting to a future operational environment, but confounding makes such generalization uncertain, if not impossible.

One of the strategies test managers routinely use to compensate for training deficiencies is to script test player participation in test events. Scripting generally takes one of two forms. In the first case, test players simply are told what to do and when to take those actions—operator participation by the numbers, so to speak. A second form of scripting is to "train up" test players using the same or similar scenarios used later during actual test runs. In either case, the outcome is similar. Test player performance variation in response to operational cues during actual test runs is reduced or virtually eliminated. This reduction in variance makes it impossible to relate user performance to test outcomes. Scripting is an insidious approach to handing test player training deficiencies because it provides a superficial appearance of operational validity, but actual test results are of little more value than those obtained in a developmental or technical test. The hardware component of the manned system is responsible for all of the observed performance variation. In effect, scripting undermines the rationale for progressing from technical or developmental testing to *operational* testing using a manned system. The operators' contribution to overall system performance cannot be determined.

When faced with this argument for better up-front preparation of test players, advocates of traditional testing practices frequently reply, "But isn't some data better than none?" The implication here is that obtaining some data on system capabilities, even if those data might be badly flawed, is better than getting none. At one level, this argument is difficult to refute. Having some test data, particularly if those results support an argument for system effectiveness, does provide a security blanket, of sorts. The problem is one of downstream risk. Certainty as to whether an observed test outcome might occur in a future combat situation is substantially reduced. In brief, the likelihood for unpleasant surprises is high.

Pretest training is harder now

Over the past several decades, the training–testing problem has been aggravated by the kinds of systems being developed and fielded. Advances in information technology have speeded the deployment of a class of systems that are termed "knowledge-based" (Dekker 2002). High levels of user skill and knowledge are necessary for the successful employment of this class of systems. Knowledge-based systems raise user skill levels because they shift the focus of operator performance away from what are termed skill- and rule-based performances (e.g., operating equipment and following procedures) and emphasize knowledge-based performances. In such systems, cognitive, knowledge-based operator performances such as planning, problem solving, and critical thinking are key user performance requirements. Moreover, many features built into the "hard" system are included to support users in performing these critical functions.

Following an assessment of training requirements for future conflicts, the Defense Science Board (DSB) observed that "Current training does not prepare individuals or units for new, dynamic cognitive demands" (DSB 2003). Similarly, in the wake of the fratricides committed by Patriot air and missile defense units during the major combat operations phase of

Operation Iraqi Freedom (OIF), the board of inquiry looking into those incidents criticized Patriot training for emphasizing "rote drills" versus the "exercise of high-level judgment." In our post-OIF assessment of the Patriot fratricides performed at the invitation of the commanding general of the Army's air defense center, we concurred that much current Army training stresses skill- and rule-based performances, but does not adequately address knowledge-based performance requirements (Hawley and Mares 2006).

To sum up, knowledge-based systems place a premium on user expertise. Expertise is a function of users' knowledge, skill, and job-relevant experience. Moreover, expertise, as the term is normally used, takes time and the right kinds of on-the-job experiences to develop. Norman (1993) asserts, for example, that for any complex activity, a minimum of 5,000 hours of training is required to turn a beginner into a journeyman-level performer. Training of the sort necessary to develop essential system expertise is termed deliberate practice and consists of relevant (skill-focused) practice with expert feedback (Ericsson and Charness 1994). In our observation, most of today's military training does not meet the definition of deliberate practice.

With respect to training and operational testing, the bottom line on the present discussion is clear: If we are dealing with a complex, knowledge-based system like Patriot or many of the battle command systems coming into the Army inventory, essential levels of user expertise cannot be developed following standard pretest train-up practices. Traditional new equipment training (NET) focusing primarily on equipment operation (skill-based performances) followed by a relatively short period of unit training emphasizing employment procedures (rule-based performances) will not produce test players capable of demonstrating critical system capabilities. With respect to pretest Patriot training, the Department of Defense's (DoD) Director of Operational Test and Evaluation (DOTE) summarized the situation quite well as follows (DOTE 2008):

> *"The level of expertise required for PAC-3 (Patriot Advanced Capability–Three) PDB-6 (Post-Deployment Build 6) operations exceeds the current Army training standard. … The operational impact of [these deficiencies] includes less robust and less effective defense of critical assets, an increased probability that operator error will lead to not engaging threatening targets and/or engaging friendly targets, and longer downtimes when reliability failures occur."*

More than the Army's training standard is at issue here, and DOTE's statement can be generalized to other systems as well. Our experiences support the DSB's observation that current training methods generally are inadequate for the knowledge-based performance demands inherent in many of today's systems. If the test community wants to avoid increasing criticism about hollow testing, this reality must be reflected in pretest training practices. The next section begins our discussion on how this might be accomplished.

Pretest training solutions: The "gold" standard

The first issue to be addressed in resolving the training–operational testing impasse is, "What would a satisfactory pretest training program look like, and how would it be carried out?" We refer to this satisfactory situation as the gold standard. The first step in meeting the gold standard involves the selection of a test organization, or unit. This unit would have met four preliminary conditions: (1) fully manned, (2) all test players qualified in the appropriate military occupational specialty (MOS), (3) all participating personnel fully trained and verified on any predecessor systems (little or no performance remediation required), and (4) all personnel stabilized for the period of pretest train-up and operational testing.

A second precondition for meeting the gold standard is achieving a stable performance setting prior to the start of testing—and preferably before the start of unit training. A stable performance setting means that (1) any equipment (hardware and software) involved in the test is sufficiently mature to support reasonably uninterrupted use and (2) the doctrine and operational procedures characterizing the system's employment have been developed and subjected to preliminary validation. Equipment and procedural documentation must also have been produced and made available to test players for training and follow-on reference.

The third gold standard requirement is the conduct of pretest training itself. Per Army Regulation (AR) 73-1, test players participating in operational tests must be trained to "deployment standard." Enforcing this operator and crew performance standard is essential if test results are to be considered valid vis-à-vis generalization of these results to future combat operations. Training to deployment standard will involve (1) adequate NET, or *orientation* to any new equipment (hardware and software) coming to test; (2) adequate follow-on unit training on the system coming to test; (3) training in usage concepts (doctrine and tactics training); (4) time to develop required operational proficiency—hands-on training to the level required by the test; and (5) test player performance verification prior to the test, with provisions for training remediation when proficiency goals are not met.

From a testing theory perspective, it is difficult to argue with these. If we want to conduct valid testing and escape escalating outside scrutiny and criticism, these steps should be taken. However, anyone familiar with operational testing in today's environment has to conclude that the requirements underlying the gold standard are rarely met, and may be increasingly unachievable for reasons discussed previously. Based on our experience in recent Army tests, we fall short of meeting the gold standard for any of the listed reasons:

1. The unit stabilization period is too short given the complexity of many new, knowledge-based systems—if stabilization is ever actually achieved.
2. Training requirements are undefined or remain a moving target.
3. Pretest training conduct is methodologically inadequate or inappropriate.
4. Equipment schedule slips encroach on planned training time.
5. Failure to achieve a stable performance setting prior to the onset of the test or to support unit training.
6. Equipment cost overruns are paid for out of planned training funds.

Even when an honest attempt is made to meet the requirements underlying the gold standard prior to testing, we often end up not getting the job done. During the period leading up to the test, the factors listed come into play, singly and in combination, and begin to degrade the basis for valid testing. Eventually, the situation is degraded to the point mentioned earlier where what started out with all good intentions and planning as an operational test is reduced to little more than a technical demonstration. We observed a situation somewhat like this during the year-long run-up to the operational test for Patriot PDB-6 in the aftermath of the OIF fratricides. Our pretest training readiness rating for the PDB-6 test was "green/red." The rating was green in the sense that all of the training "events" planned for the pretest train-up period had been completed, but individual and crew proficiency objectives were not achieved (i.e., red). We concluded that in spite of the year-long train-up period, test players were not able to perform at the level required by the test's objectives. The impact of this failure to meet test player proficiency objectives is summarized in the DOTE's comment on the PDB-6 operational test cited previously. Clearly, a new view of testing for complex, knowledge-intensive systems is required. The next section outlines several of our thoughts on how to break out of the worsening impasse between pretest training and valid operational testing.

A modest proposal for change

Many of the military systems currently under development are software dominated in the sense that key system capabilities are resident in software, as opposed to the mechanical or electronic components of the system. Patriot PDB-6 certainly falls into this category, as do many of today's battle command systems. That being the case, a good starting point for considering potential changes to operational testing concepts is to look at the way commercial software companies test and deploy their products.

Although practices vary, the software release cycle proceeds roughly as follows. To initiate the formal cycle, an initial version of the software generally considered complete is evaluated in an alpha test inside the organization developing the software. Alpha testing is performed by personnel other than the software's developers. When alpha testers are satisfied that the system works the way it is supposed to work (e.g., If I press "A" on the keyboard, the display shows an "A."), it is released to a limited number of users, external to the organization, who then conduct a beta test. During the beta test, the software undergoes extensive usability testing and operational debugging. Under standard industry practices, beta testers typically use the system for a "soak period," generally about 1 year, and then human factors engineers go out and interview the participants to find out what is working, what is not, and what changes are required (Savage-Knepshield 2009). The intent is to provide developers with constructive, actionable feedback on usability problems, bugs, and other flaws. Both alpha and beta testing frequently are done iteratively with prospective end users over the course of the system's design and development process.

The beta version of the software is the first version released outside the organization or community that developed it, and it is done to support evaluation in a "real-world" setting. Upon completion of the beta test, the software is considered a "releasable candidate" in that its quality is considered sufficient for more general distribution. In most situations, the software evaluation process does not end with the beta test. Most commercial organizations monitor products for latent bugs and other problems long after the product is sold to the general public.

The beta testers themselves are considered crucial to the evaluation process. In many situations, beta testers are customers or prospective customers carefully selected to provide a "user test" for the software product. The beta testers' job is to provide sophisticated feedback to developers on software usability, bugs, flaws, and other problems. Beta testers are able to do this because of their knowledge of software testing methods and the application domain. They understand how the software works, generally, and what it is supposed to do. In this sense, software beta testers are analogous to the test pilots who first fly aircraft prototypes. Test pilots are carefully selected, highly trained and experienced in flight operations, trained in flight testing methodology, and familiar with aeronautical engineering concepts. They understand how and why advanced aircraft are tested. Because of their extensive background and experience, they are able to give the prototype a thorough shakedown before it is released to less experienced users.

What does this discussion have to do with military operational testing? How might the software testing model be applied to military test and evaluation? The parallels are rather direct, in our view. To begin, the software beta test notion applies most directly to the system shakedown and refinement period between the end of developmental or technical testing (i.e., alpha testing) and deployment of the system to first operational units equipped. Early-on user tests would be viewed as analogous to software beta tests and would be approached in much the same manner. Their objective would be to provide in-depth, real-world feedback to the system's developers. User test participants like beta testers would be users or prospective users of the eventual product. However, beta test players would not be selected in the same manner as today's test players. Instead, they would be selected more like commercial software beta testers or aviation test pilots. They would be experienced in the technical and tactical domains involved and trained in test and evaluation methodology. Their job would be to provide meaningful feedback to software and concept developers on the prospective system's fitness for use and the validity of proposed usage concepts. Such feedback cannot be obtained from less sophisticated, rank-and-file test players. From our experience, typical test players "don't know what they don't know."

An example from a previous Army test will illustrate the potential utility of the testing concept we are advocating. Around 15 years ago, during the limited user test of a biological detection system, the test player population consisted of Army crews selected and trained in the usual fashion along with several control crews consisting of Ph.D.-level technical personnel from one of the Army's biological laboratories. As the test progressed, test management personnel observed that the system was not performing as intended. Rates of proper agent identifications were not satisfactory: misses were frequent, false alarm rates were high, and equipment malfunctions were common. Their immediate reaction to these system problems was to blame test player training as inadequate. A second proposed explanation was that the MOS involved was not appropriate for the bio detection job. A MOS with a higher aptitude composite and more intense background training in the underlying content domain might be required.

After reviewing preliminary test data, the human factors team (of which the lead author was a part) pointed out that the performance of the control crews was not statistically better than that of the military crews. The control crews did not perform any better than the Army crews participating in the test, and the control personnel were as good as it gets, so to speak. After reviewing these results, a decision was made to interview the control crew members to find out why their performance using the proposed bio detection system was inadequate. Because of their extensive background in the technical domain involved, control crew members were able to tell us why the test results were as they were. The explanation had nothing to do with test player training or the MOS involved. It had to do with technical inadequacies of various components of the detection system, and these problems could not be eliminated through training or personnel solutions. A different system concept was required.

Having a sophisticated group of what might be termed beta testers participating in the test provided information that would not have been available using routine methods for test player selection and training. The insight provided by the control crew members saved the Army from going down an inappropriate remedial path and speeded the development of a biological detection capability that was able to meet expected performance requirements.

Use of something similar to the beta testing concept during early and midrange user testing would avoid the expensive, "cast-of-thousands" exercises that are now common in operational testing. Fewer personnel would be required, train-up times for participants would be considerably shorter, and posttraining results would be more satisfactory. In our view, similar or better test results could be obtained far more cost-effectively using the beta testing concept. However, abandoning traditional concepts for test player selection and training in favor of sole use of control test participants might be a stretch for many in the testing community. There also might be significant problems with the current regulatory structure governing operational testing (e.g., the need for "representative" users as test participants). However, including carefully selected beta test or control crews in the test player population would be useful in providing improved feedback to system and concept developers. Data from control test players might also be useful when interpreting results from routine test participants, as was the case with the biological detection system.

Summing up

Adequate test player training is essential to effective operational testing of complex human–machine systems. In the absence of adequate test player training, test validity is compromised and generalizing test results to future operational settings is risky. Historical pretest train-up problems are being exacerbated by the development and fielding of complex, knowledge-based systems used in complex ways. The end result of these developments is that providing suitably trained test players following old testing practices is increasingly unachievable. Moreover, persisting in old test player preparation practices inevitably results in hollow tests that do not provide information essential for system and concept evaluation and refinement, as well as inviting increasing skepticism and criticism concerning the validity of test results. New concepts and methods for operational testing, particularly early user testing, clearly are required.

We have proposed several modified operational testing practices that hold promise for mitigating the growing impasse between pretest training practices and valid operational testing. These modified practices include (1) adopting something similar to the software beta testing concept for early operational tests, and (2) increased use of control test players. The rationale for these proposed changes in operational testing concepts and procedures is straightforward: Evaluators must obtain valid and insightful data concerning the subject system's performance potential and limitations along with feedback on the efficacy of proposed usage concepts. Inexperienced test players cannot provide this essential feedback. Using data from inexperienced and often inadequately trained test players as the sole basis for system evaluation and acquisition decision making increases the risk that systems will be fielded without an adequate assessment of their fitness for later operational use. ❑

John K. Hawley, Ph.D., is chief of the U.S. Army Research Laboratory's Human Research and Engineering Field Element at Fort Bliss, Texas. Since receiving his doctorate, Dr. Hawley has worked as an applied psychologist for more than 25 years in a variety of government and private-sector organizations. He recently served as project lead for an Army effort to examine human performance contributors to fratricides involving the Patriot air and missile defense system during the Second Gulf War and recommend potential solutions. Dr. Hawley is now working with the air defense community to implement and evaluate selected recommendations involving human-systems integration practices, test and evaluation methods, personnel assignment practices, and operator and crew training. The primary thread running through Dr. Hawley's professional experience is helping people and organizations manage the human side of transitions to new systems, processes, and technologies. E-mail: jhawley@arl.army.mil

Anna Lucia Mares received a bachelor of science in electrical engineering from the University of Texas at El Paso. She has worked for the U.S. Army Research Laboratory for more then 15 years, which has included development, analysis, and evaluation of numerous joint and Army Programs. Initially her efforts focused on ground, soldier, and munitions systems with more recent involvement centered on the human performance dimension of the Air and Missile Defense systems. In 2007, Ms. Mares received the prestigious Army Research Laboratory Award for Analysis for her analytical contributions in identifying significant issues involving the Patriot Air and Missile Defense system usage during Operation Iraqi Freedom. Her efforts resulted in recommendations for material and training solutions that increased friendly protection involving the Patriot air and missile defense system. E-mail: Anna.Mares@us.army.mil

References

DSB (Defense Science Board). 2003. *Training for future conflicts*. Washington, DC: Office of the Undersecretary of Defense for Acquisition, Technology, and Logistics.

Dekker, S. 2002. *The field guide to human error investigations*. Burlington, Vermont: Ashgate.

DOTE (Director Operational Test and Evaluation). 2008. *FY 2007 assessment of the Ballistic Missile Defense System (BDMS) Part I*. Washington, DC: Office of the Secretary of Defense.

Ericsson, K. A. and Charness, N. 1994. Expert performance: Its structure and acquisition. *American Psychologist*. 49 (8): 727–747.

General Accounting Office (formerly). 2000. *Best practices: A more constructive test approach is key to better weapon system outcomes* (GAO/NSAID-00-199). Washington, DC: U.S. Government Accountability Office.

Hawley, J. K. 2007. Training and testing: A complex and uneasy relationship. *The ITEA Journal of Test and Evaluation*. December 2006/January 2007: 34–40.

Hawley, J. K. and Mares, A. L. 2006. *Developing effective supervisory control for air and missile defense systems* (ARL-TR-3742). Adelphi, MD: U.S. Army Research Laboratory.

Norman, D. A. 2002. *The design of everyday things*. New York: Basic Books.

Savage-Knepshield, P. A. 2009. Personal communication, February 8, 2009.